HARVARD UNIVERSITY
A. ALFRED TAUBMAN CENTER FOR STATE AND LOCAL GOVERNMENT
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

TAUBMAN CENTER POLICY BRIEFS

March 2008

Searching the Family Tree for Suspects:

Ethical and Implementation Issues in the Familial Searching of DNA Databases

By David Lazer*, Program on Networked Governance, Kennedy School of Government*

Familial searching – that is, the use of DNA data from known individuals (typically convicted felons) to identify *relatives of those individuals* as potential suspects – has the potential to increase greatly the number of criminal investigative leads produced by existing DNA databases in the United States. The first documented use of a familial search of a DNA database in the US, the Sykes case, offers a dramatic glimpse of the power of this technique. Deborah Sykes, of Winston-Salem, NC, was raped and murdered on her way to work on August 10, 1984. The investigation that followed led to the conviction of Darryl Hunt for her murder. Though a subsequent test of the DNA from the crime scene showed no match with Hunt's DNA, the courts rejected this finding as definitive proof of his innocence. In 2003, the DNA profile from the crime scene evidence was run against the state database. While there was no direct match, there was a near match to the crime scene evidence from a profile in the offender database, suggestive of a potential familial relationship between the offender in the database and the source of the crime scene sample. This hypothesized familial link eventually led the police to Willard Brown, who confessed to the crime. Brown was subsequently convicted, and Hunt, who had served eighteen years of his life sentence, was released from prison.

The investigative power of familial searching rests on a combination of sociological and scientific foundations. Sociological research suggests that there is a strong tendency for criminal behavior to cluster in families. As a result, the near relatives of those in DNA databases are at a relatively high risk to commit a crime. Scientifically, familial searching is possible because the DNA profiles of related individuals tend to be similar in statistically predictable ways. The practical statistical question is whether the DNA of related individuals tends to be similar enough so as to distinguish a relative out of the very large number of non-relatives in a DNA database.

The answer to that question is, at least in some cases, a clear yes. The UK has aggressively pursued familial searching in hundreds of cases, with some dramatic success stories, and Bieber, Brenner, and Lazer (2006) demonstrate the potential effectiveness of familial searching on a much broader scale. The US has not nearly as aggressively pursued the use of familial searching as the UK —even though the database in the US is better suited for familial searching. (The

David Lazer

David Lazer is Director of the Program on Networked Governance at the Kennedy School of Government, and Associate Professor of Public Policy. He is editor of DNA and the Criminal Justice System: The Technology of Justice *(MIT, 2004). See http://www.hks.harvard.edu/lazer.*

A. Alfred Taubman Center for State and Local Government

The Taubman Center and its affiliated institutes and programs are the Kennedy School of Government's focal point for activities that address urban policy, state and local governance and intergovernmental relations.

Taubman Center Policy Briefs are short overviews of new and notable research on these issues by scholars affiliated with the Center.

The contents reflect the views of the author (who is responsible for the facts and accuracy of the material herein) and do not represent the official views or policies of the Taubman Center for State and Local Government.

*A. Alfred Taubman Center for State
and Local Government
John F. Kennedy School of Government,
79 JFK St., Cambridge, MA 02138*

***Telephone:** (617) 495-2199*
***Email:** taubman@ksg.harvard.edu*
www.ksg.harvard.edu/taubmancenter

US database includes more genetic data on each individual, making it easier to distinguish relatives from non-relatives).

Bieber, Brenner, and Lazer conservatively estimate that familial searching could increase the number of investigative leads produced by the DNA database system by 40%. Given that the total number of investigations aided from the database system to date in the US exceeds 60,000, it is plausible that the widespread use of familial searching could produce many thousands of useful leads almost overnight, just based on the data already in the offender and crime scene databases.

Such an application of these databases, however, is not what was originally envisioned by the legislatures that authorized their creation. Familial searching would effectively incorporate into the database millions of individuals who have never even been suspected of a crime, raising significant ethical and policy issues. This policy brief explores the practical ethical and policy choices facing federal and state governments regarding whether and how to implement familial searching. The implementation issues are explored first, and the policy issues (in light of potential implementation) are explored second.

Implementation issues

Familial searching involves searching a large database of known sources for a potential relative of the source of crime scene evidence. This is, essentially, a challenge of finding the needle in the haystack, with the key problem being the separation of the necessarily few (potentially zero) relatives in the database from the many non-relatives.

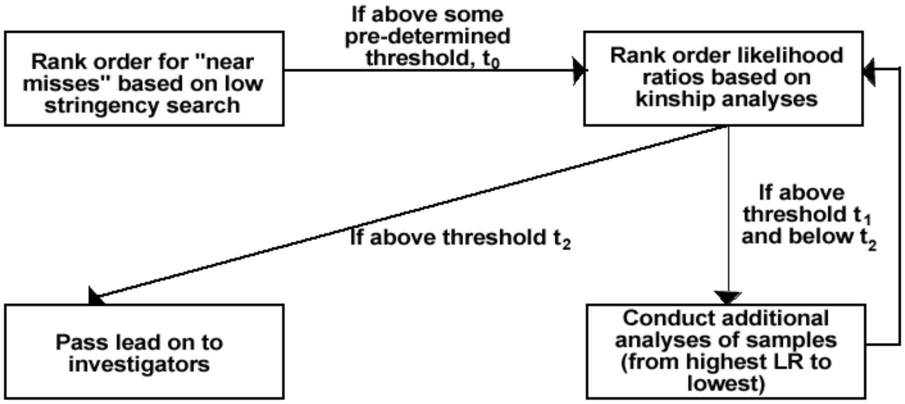
A familial search would need to take place within the overarching architecture of CODIS (Combined DNA Index System), the US software/hardware system that organizes and distributes genetic data for the criminal justice system. The most efficient way to conduct a familial search is to calculate kinship likelihood ratios, with higher likelihood ratios indicating a higher chance of a familial relationship. However, the current CODIS architecture does not allow for direct application of kinship analysis.

Until a fully functioning kinship module is added to CODIS, alternative solutions must be found. In the short run, the way that familial searching would need to be implemented is to use current CODIS functionality to search for a minimum threshold number of matched alleles (genetic markers) between crime scene data and known sources in the database (the more shared alleles, the more likely a familial relationship). Kinship analysis can then be performed manually on this more limited set of profiles, where likelihood ratios above some threshold would be reported to investigators as a "familial hit" or selected for further analysis. Further analysis of the offender and crime scene sample, while potentially costly, would greatly reduce false leads. For example, examination of Y chromosome profiles (where the Y chromosome is passed unchanged from father to son) could eliminate well over 99% of false leads. In the moderate run, labs can use the kinship analysis functionality that is built into the missing persons database system (CODISmp) to produce a rank ordering of likelihood ratios. This would eliminate the step of identifying potential leads through shared alleles.

There is no "one size fits all" familial search policy that will be desirable for all crime labs and all cases.

When investigators are provided a familial "hit" (i.e., the name of the individual in the database who is believed to be related to the potential perpetrator) their first step will be

Figure 1: Decision Sequence For Familial Searching


Rank order for "near misses" based on low stringency search
If above some pre-determined threshold, t_0
Rank order likelihood ratios based on kinship analyses
If above threshold t_2
If above threshold t_1 and below t_2
Pass lead on to investigators
Conduct additional analyses of samples (from highest LR to lowest)

to examine genealogical data to determine if that individual has any near relatives who are potential candidate sources of the crime scene DNA.

For example, if the hypothesized relationship is sibling, does the individual have a brother of an appropriate age (e.g., non-infant)? Further, does that brother live near the crime scene? These facts would need to be evaluated before further costly and intrusive investigation. In the long run, kinship analysis could be combined with genealogical and geographical data to automatically produce likelihood ratios for the combined data.

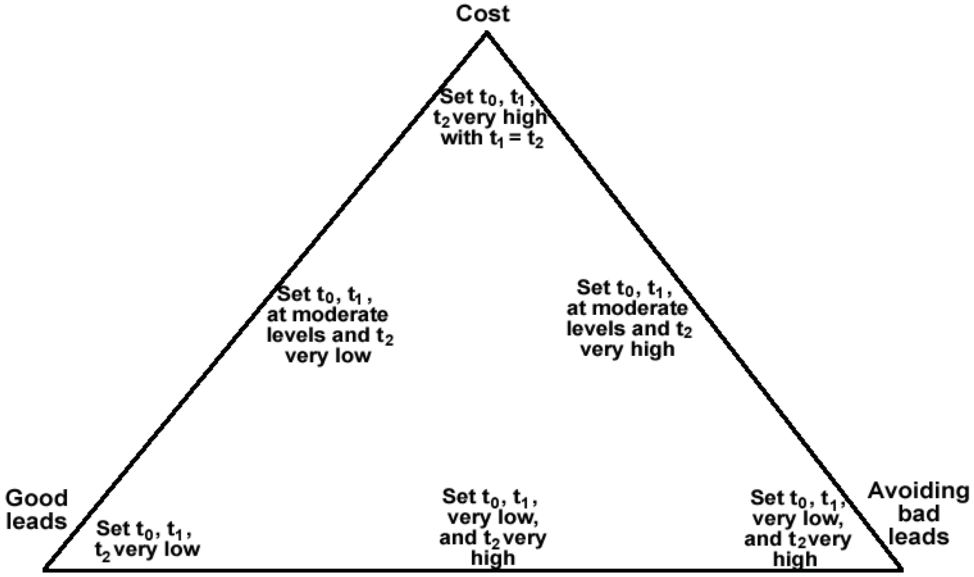
An alternative way to implement familial searching would be to use a decision rule based solely on allele sharing. Thus, for example, some states require a certain number of shared alleles between samples. (Some states, for example, have a “20 allele rule,” requiring at least a single allele match for all 13 loci — there are two alleles at each locus — and a minimum of 20 alleles to match out of the possible 26.) Such a decision rule, however, is grossly inefficient because it does not take into account that a shared allele that is rarely observed is a much more powerful signal of a family relationship than one that is frequently observed. An approach incorporating kinship analysis could easily produce fewer false positives and more true positives.

Figure 1 summarizes the sequence of steps in familial searching. The key decision parameters are: (1) the thresholds used in evaluating the number of shared alleles between the crime scene profile and the profiles in the database, (2) the likelihood ratio thresholds used to decide whether to conduct additional analysis and/or pass a lead on to investigators, and (3) an assessment of how much additional analysis to conduct on samples.

There is no “one size fits all” familial search policy that will be desirable for all crime labs and all cases. Exactly what choices a lab must make on these three parameters depends on the answers to three questions:

- How much are you willing to spend on familial searching?
- How many potential perpetrators do you want to identify with familial searching (true positives)?
- How many wrong doors are you willing to knock on (false positives)?

Figure 2: Decision Heuristic for Implementation of Familial Searching


Cost
Set t_0, t_1, t_2 very high with $t_1 = t_2$
Set t_0, t_1, at moderate levels and t_2 very low
Set t_0, t_1, at moderate levels and t_2 very high
Good leads
Set t_0, t_1, t_2 very low
Set t_0, t_1, very low, and t_2 very high
Set t_0, t_1, very low, and t_2 very high
Avoiding bad leads

Obviously, in the ideal world, familial searching should be cheap, and yield only good leads, but no bad leads. However, these three goals are (at present) in conflict, and how one implements familial searching depends on their relative priority. Consider three scenarios for the relative priority of these three goals:

Keeping cost low and avoidance of bad leads are the priority goals: In this case laboratories would choose a somewhat *higher* allele-match threshold, as well as a very high likelihood-ratios threshold at which to pass a lead on to investigators, and conduct little or no additional analyses on samples. This would be quite cheap to implement (essentially, analyzing existing data), and would yield very few bad leads. The downside is that many potentially useful leads would be ignored.

Identification of good leads and avoidance of bad are the key priority goals: In this scenario, the laboratory would choose a relatively *low* allele-match threshold and a relatively low likelihood-ratio threshold at which to conduct extensive additional analyses of the relevant samples (presumably, prioritized in order of highest ratio to lowest) and a high likelihood-ration threshold before reporting a familial hit to investigators. If cost were truly no object, it should be possible to greatly lower the probability of a false lead and increase the likelihood of including any true first-degree relative to a high probability.

Keeping cost low and identification of good leads are the priority goals: In this case, the laboratory should set relatively low allele-match and likelihood-ratio thresholds at which to pass on the information to law enforcement, and conduct little or no additional analysis of the samples. Of course, this passes on costs to investigators, who must pursue more long shot leads.

The relative priority of these three goals should vary with the context. The willingness to expend resources and knock on wrong doors should be higher in a serial murder case than a burglary, for example. Figure 2 offers a decision triangle on which to base policy. The closer a laboratory places itself to a particular corner, the higher that priority.

Federal versus state implementation of familial searching

In the short run, familial searching would need to be implemented almost exclusively at the state level. The reason for this is twofold: (1) the federal system is a much larger haystack, making it harder to identify relatives, and (2) it is likely that relatives of offenders live in the same state as the offender. (Thus, it is less likely the needle would be in the federal haystack.) At the present time, searches at the federal level would yield many profiles that look like potential near relatives just by chance.

In the longer run, it would be simple to incorporate kinship analysis into the federal database, using variable likelihood thresholds depending on the proximity of other states. And in high priority cases, one could lower those thresholds and conduct further analyses of those samples. Further, if additional data were automatically extracted from samples, more definitive searches of the federal databases would be possible (see below).

Retention of samples and familial searching

One of the key controversies about the DNA databank system is the retention of physical samples from known offenders. On one hand, the digitized profiles in the computer databases represent only a tiny fraction of the information that is possible to extract from those samples. On the other, the retention of samples from offenders raises significant privacy issues, because, for example, sensitive medically relevant information can still be extracted from retained physical samples. This has led some to advocate destruction of the physical samples from the offender databank, which no jurisdiction currently does. Familial searching, as discussed above, would be more effective if samples were retained, and additional data were extracted.

It is likely that the database system will begin incorporating additional information from samples at some point in the future. There are supply and demand side reasons for this. On the supply side, the marginal costs of analyzing samples for additional information is dropping precipitously. In the not too distant future, it will be possible to extract magnitudes more genetic information than is currently done —at essentially zero additional cost.

Familial searching would be more effective if offender samples were retained, and additional data were extracted.

On the demand side, there is some benefit to the extraction of additional information, beyond increasing the effectiveness of familial searching. For the current objective of the CODIS system (linking samples from the qualifying offender database to unknown crime scene samples, and linking unknown crime scene samples to each other), the current amount of data that is extracted is more than enough, *if those samples are pristine*. However, crime scene samples are often degraded, or mixtures of multiple sources of DNA. In such cases, the current system may not offer enough statistical power to identify suspects who are in the database. Additional genetic data would be quite useful in such scenarios. The incorporation of large amounts of additional genetic data would also prove quite useful for familial searching, essentially eliminating the cost/good lead/bad lead trade-offs discussed above. Given orders of magnitude more genetic data, it will be possible to identify first-degree relatives with high precision, even in a search of the federal database. (Although the dilemma may then be pushed further out into the family tree.)

Ethical and policy concerns

The ethical concerns about the use of familial searching center around the balance between the benefits of using an existing resource to produce useful leads for investigations, and the potential adverse effects on individuals, groups, and society. At the individual level, the concern is that one is creating a new class of individuals who are under life-long genetic surveillance. If part of the justification of putting convicts under certain types of life-long surveillance is that they broke the social contract that binds them to society, what is the justification for placing family members of convicts under such surveillance? Alternatively, given the scenario of the serial murderer, could one imagine *not* examining whether a familial search of the database would produce a useful lead?

Below is a critical examination of some of the key arguments regarding familial searching from proponents and opponents.

1) Familial searching will lead to knocking on lots of wrong doors.

As discussed above, this actually depends on how familial searching is implemented. One can implement familial searching in a fashion that minimizes knocking on wrong doors, albeit at some cost and some missed opportunities to knock on right doors. Further, all investigative techniques pose the risk of knocking on wrong doors on occasion. The only way to truly eliminate the risk of knocking on wrong doors is to eliminate all police investigations.

However, the knocking on wrong doors (and sometimes even right doors) does pose a unique risk in the case of familial searching, which is the unintentional revealing of genetic (non)relationships. The rate of non-paternity in the US is commonly estimated to be in the 5-10% range. It is inevitable that investigations will therefore sometimes discover, and potentially reveal, cases of nonpaternity. Should familial searching be conducted, special training of investigators to use genetic information should therefore be required, so as to avoid causing harm to individuals in the course of an investigation.

It is inevitable that investigations will sometimes discover, and potentially reveal, cases of nonpaternity.

2) Familial searching will be a drain on a system that is short on resources.

This is also an implementation choice. In the short run, familial searching can be implemented in a fashion that poses very little drain on DNA laboratories. Further, the identification of useful leads potentially will reduce the strain on investigative resources. In the long run, it should be possible to implement familial searching in a fashion that imposes virtually zero cost on the laboratories.

3) Familial searching would disproportionately incorporate minorities into the database.

This is true. African Americans, for example, suffer from conviction rates for felonies approximately seven times that of Caucasians. However, it is ambiguous what the implications of this observation are for the application of familial searching.

One issue is that to the extent that these uneven numbers reflect a series of biases in the criminal justice system (e.g., from statutes or biased police practices) familial searching (along with many other practices in the criminal justice system) would amplify those inequities. However, minority communities would also likely be the disproportionate beneficiaries of familial searching for two reasons. First, crime occurs disproportionately within racial and ethnic groups. Victims of crimes that might be investigated through familial searching will thus more likely be from minority communities. Second, minorities are disproportionately victims of violent crime. The public safety benefits of familial searching would therefore disproportionately accrue to minority communities. This does not obviate concerns about higher levels of surveillance of minority communities, but it suggests that the reality of the costs and benefits to minority communities is complex.

4) There is nothing to fear from being in the database (either directly or indirectly) as long as you do not commit a crime.

This assertion rests on a false confidence in the infallibility of DNA in identifying the perpetrators of crime. The power of DNA rests firmly in the arms of a necessarily fallible criminal justice system. How could DNA incorrectly lead to the conviction of an innocent person? The first, and fundamental, issue is that the match between an individual and a sample at the crime scene does not indicate how the sample arrived at the crime scene. DNA does not come with a time stamp or delivery receipt. In short, the match may be correct, but the narrative around the match could be wrong. Second, there is always the possibility of human error in the examination of the evidence.

These observations are not meant as a critique of the use of DNA in the criminal justice system, because, in the absence of divine intervention, there is no infallible way to identify perpetrators of crime. However, to be "in the system" comes at the cost of some small Hitchcockian probability of having one's DNA in the wrong place at the wrong time.

Can we, as a society, tolerate the virtual certainty that some innocent person will have their DNA in the wrong place at the wrong time and be identified due to familial searching? This is, of course, possible in the absence of familial searching (or DNA altogether), but it is part of the price *particular* individuals pay for being under enhanced surveillance. If this is a tolerable risk in the case of familial searching, is there a reason (other than cost) not to incorporate everyone into a universal database, so that risk can be equitably distributed?

More generally, familial searching as a practice raises a series of civil libertarian concerns about giving informational power to the government, around potential misuse of the data. Genetic data are particularly sensitive information, and the question arises how to limit their potential misuse. There are multiple levels of potential safeguards. The most stringent involve limiting the information that government retains, i.e., making certain types of abuses impossible in the short run.

To be "in the system" comes at the cost of some small Hitchcockian probability of having one's DNA in the wrong place at the wrong time.

The downside of this approach is that it may limit the capacity of the government to achieve legitimate ends as well. For example, the destruction of samples would make it impossible for the government to inappropriately extract additional data from those samples. However, it would also eliminate the possibility of extracting additional information so as to aid an investigation legitimately. Other potential safeguards involve procedural limits in how to use information, and transparency in the utilization of information. Such safeguards have the advantage of allowing reuse of information for legitimate purposes, but the disadvantage that they do not pose an absolute barrier to abuse. (For further discussion on the alternative mechanisms for limiting the informational power of the government *vis a vis* DNA, see Lazer and Mayer-Schönberger below).

RELATED PUBLICATIONS

Frederick Bieber, Charles Brenner, and David Lazer, “Finding Criminals Through DNA of Their Relatives,” *Science*, June 2, 2006.

David Lazer, *DNA and the Criminal Justice System: The Technology of Justice* (MIT press, 2004).

David Lazer and Viktor Mayer-Schönberger, “Statutory Frameworks for Regulating Information Flows: Drawing Lessons for DNA Data Banks from other Government Data Systems.” *Journal of Law, Medicine, and Ethics* 34, 2006.

TAUBMAN CENTER POLICY BRIEFS

March 2008 “The Greenness of Cities,”
by Edward L. Glaeser (Harvard University) and Matthew Kahn (UCLA)

February 2008 “The Seven Big Errors of PerformanceStat,”
by Robert D. Behn (Kennedy School of Government)

PB-2007-6 “Flypaper and Fungibility: Evidence From the Master Tobacco Settlement,”
by Monica Singhal (Kennedy School of Government)

PB-2007-5 “The Rise of the Sunbelt,”
by Edward L. Glaeser (Harvard University) and Kristina Tobio (Kennedy School of Government)

PB-2007-4 “Productivity Spillovers in Health Care,”
by Amitabh Chandra (Kennedy School of Government) and Douglas O. Staiger (Dartmouth College)

PB-2007-3 “High Performance in Emergency Preparedness and Response: Disaster Type Differences”
by Herman B. “Dutch” Leonard and Arnold M. Howitt (Kennedy School of Government)

PB-2007-2 “Racial Statistics and Race-Conscious Public Policy,”
by Kim M. Williams (Kennedy School of Government)

PB-2007-1 “Transparency Policies: Two Possible Futures”
by Archon Fung, Mary Graham, David Weil, and Elena Fagotto (Kennedy School of Government)

PB-2006-2 “Beyond Katrina: Improving Disaster Response Capabilities”
by Arnold M. Howitt and Herman “Dutch” Leonard (Kennedy School of Government)

PB-2006-1 “Why are Smart Cities Getting Smarter?”
by Edward L. Glaeser (Harvard University) and Christopher Berry (University of Chicago)

PB-2005-4 “The ‘Third Way’ of Education Reform?”
by Brian Jacob (Kennedy School of Government)

PB-2005-3 “From Food to Finance: What Makes Disclosure Policies Work?”
by Archon Fung, Mary Graham, David Weil, and Elena Fagotto (Kennedy School of Government)

PB-2005-2 The Fiscal Crisis of the States: Recession, Structural Spending Gap, or Political “Disconnect”?
by Robert Behn and Elizabeth Keating (Kennedy School of Government)

PB-2005-1 “Smart Growth: Education, Skilled Workers and the Future of Cold-Weather Cities,”
by Edward L. Glaeser (Harvard University)